SARDEGNA RICERCHE

Sardegna FESR 2014/2020 - ASSE PRIORITARIO I

“RICERCA SCIENTIFICA, SVILUPPO TECNOLOGICO E INNOVAZIONE”

Azione 1.1.4 Sostegno alle attività collaborative di R&S per lo sviluppo di nuove tecnologie sostenibili, di nuovi prodotti e servizi

Progetto cluster Top Down “NIASMIC - Not Invasive Analysis of Somatic Mutations In Cancer”

Rapporto Tecnico

Specifiche tecniche del sistema per la gestione dei dati di sequenziamento, definizione dei workflow di riferimento per l’analisi dei dati, e struttura del software per l’integrazione di dati genetici e clinici.

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR SARDEGNA 2014-2020

RAPPORTO R3.4 Rapporto tecnico con le specifiche tecniche del sistema per la gestione dei dati di sequenziamento, definizione dei workflow di riferimento per l'analisi dei dati, e struttura del software per l'integrazione di dati genetici e clinici.

Il progetto NIASMIC ci pone di fronte a stimolanti sfide tecnologiche:

- necessità di strumenti di calcolo scalabili in grado di tenere il passo con una così massiva generazione di dati;
- realizzazione di una gestione efficace dell'archiviazione;
- necessità del tracciamento dei dati e della loro provenienza;
- gestione di complessi workflow di analisi;
- riproducibilità dei risultati;
- interfacce utenti semplice ed immediate per destinatari non tecnici.

In risposta a queste sfide, abbiamo lavorato per costruire un'infrastruttura automatizzata che integra strumenti open source (sviluppati internamente o disponibili pubblicamente) in un framework che consente l'acquisizione automatica dei dati di sequenziamento, la tracciatura della provenienza dei dati, la configurazione e la gestione delle pipeline personalizzate, oltre ad una virtualizzazione dei dati flessibile e affidabile.

La Fig. 1 mostra una panoramica dei principali componenti della infrastruttura, che verranno trattati in dettaglio nei prossimi paragrafi.

Sistema di tracciabilità

Tutte le attività del laboratorio NGS sono tracciate utilizzando PENELOPE, il sistema di gestione delle informazioni di laboratorio sviluppato al CRS4, che integra tre componenti:

1. Database relazionale basato su Bika-LIMS[1], potente e flessibile LIMS open-source e conforme allo standard ISO 17025;
2. Interfaccia web interamente costruita intorno ad un framework AngularJS[2];
3. RESTful API che funge da interfaccia tra il nucleo di Bika e il mondo esterno, rappresentato dall'interfaccia web e dalle applicazioni satelliti che mappano le azioni 'CRUD' (Scrittura, lettura, aggiornamento, cancellazione)

[1] https://www.bikalims.org/
[2] https://angularjs.org/

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR SARDEGNA 2014-2020

SARDEGNA RICERCHE

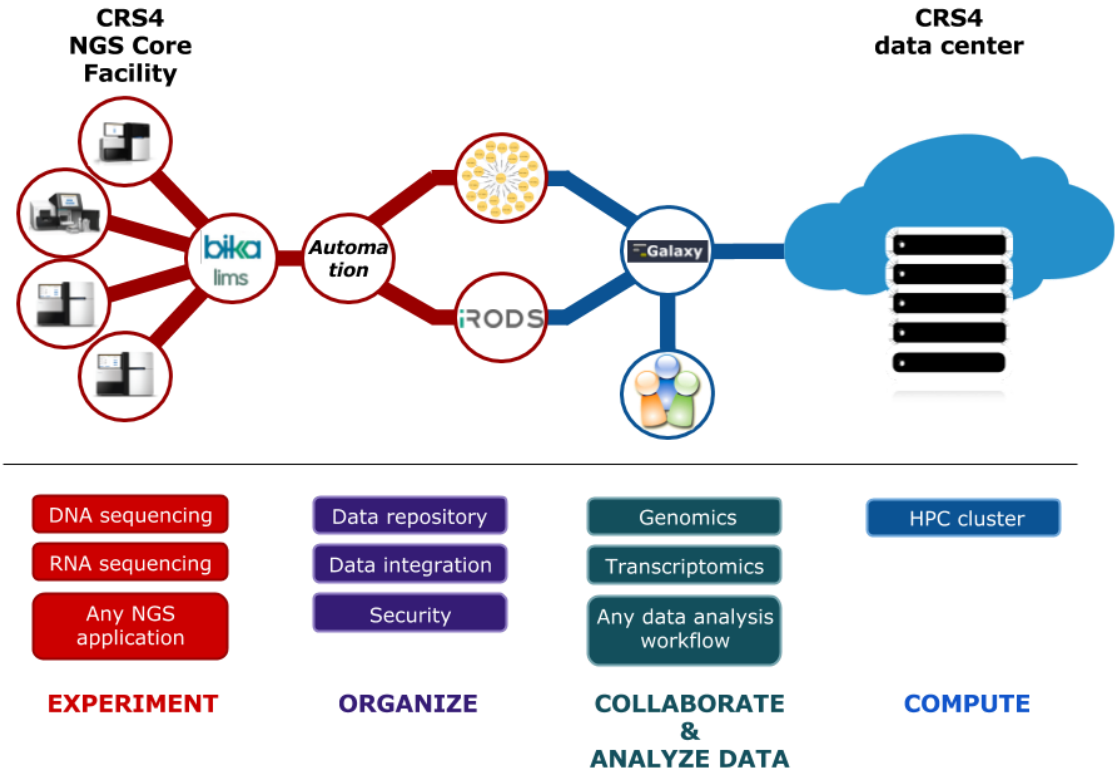

Fig. 1 - Panoramica dei principali componenti della infrastruttura CRS4: sistema di tracciabilità (**Organize**), elaborazione dei dati di sequenziamento (**Organize**), gestione dei workflow di analisi computazionale (**Collaborate e Analyze Data**), raccolta dati fenotipici (**Organize**), gestione dei dati di sequenziamento (**Organize**), pipeline di analisi (**Collaborate e Analyze Data**).

PENELOPE/Bika LIMS gestisce gli *oggetti* e gli *eventi* del laboratorio NGS. Gli *oggetti* essenziali sono i campioni in analisi, la strumentazione del laboratorio, i materiali di supporto, etc.

Gli *eventi* riguardano il cosiddetto "ciclo di vita" degli oggetti sopra citati, quindi il ciclo di vita dei campioni (vedi Fig.2 per il workflow completo).

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR
SARDEGNA 2014-2020

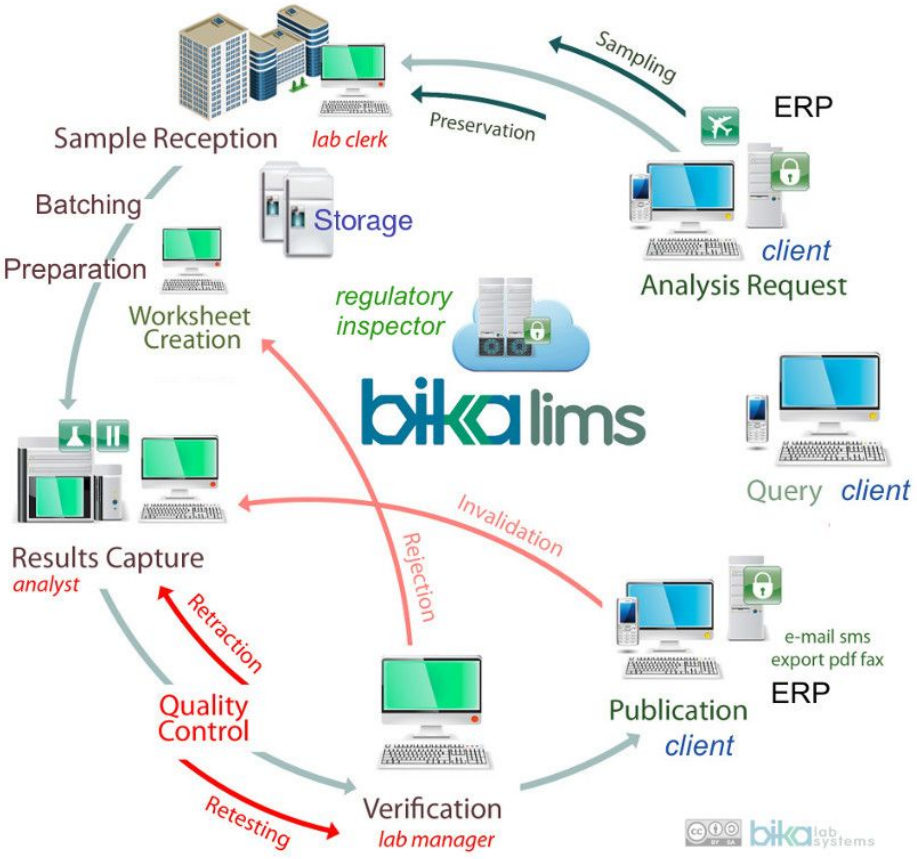

Fig. 2 - Ciclo di vita dei campioni all'interno di BIKA-LIMS

PENELOPE ci consente di essere in grado di gestire ruoli e autorizzazioni, gestire l'inventario e la catena di approvvigionamento, tracciare campioni, dare priorità alle richieste di analisi e organizzare facilmente i worksheets per il progetto NIASMIC (vedi Fig 3a e 3b)

Elaborazione dati di sequenziamento

I dati massivamente prodotti dalle macchine di sequenziamento necessitano di uno step di pre-elaborazione, per passare dal formato *raw* (specifico per ogni dispositivo sperimentale) ai formati standard gestibili da tutte le applicazioni e tool.

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR SARDEGNA 2014-2020

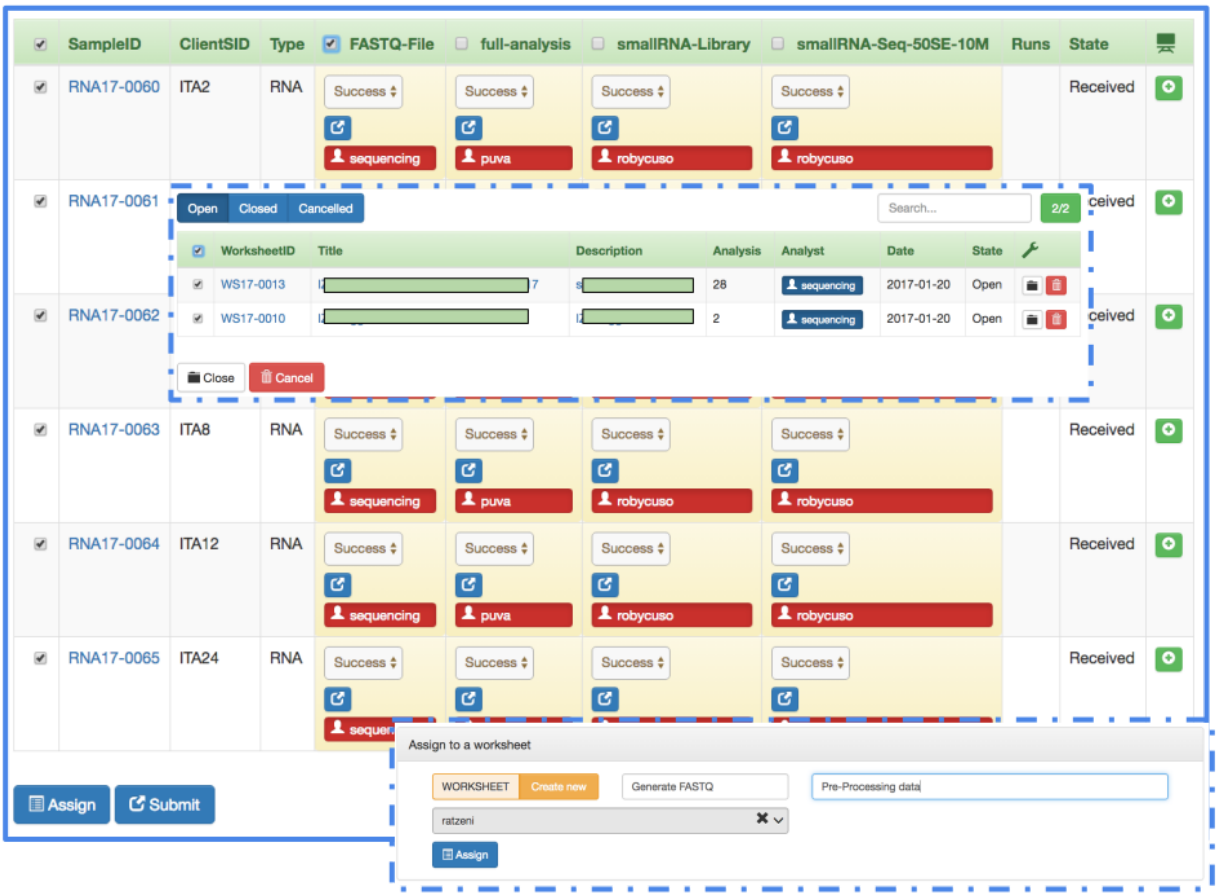

Fig. 3a - Screenshot di PENELOPE

La pre-elaborazione dei dati, la loro conversione in formato standard e l'archiviazione è automatizzata da una macchina a stati finiti, sviluppata al CRS4 e denominata PRESTA (Fig. 4).

Il nucleo di PRESTA è costruito intorno a Celery[3], un gestore open source di code di attività, basato sul passaggio di messaggi distribuiti, che cadenzano il passaggio da uno stato all'altro.

Al fine di minimizzare l'intervento umano e aumentare l'affidabilità, uno processo periodico controlla lo stato di avanzamento delle run di sequenziamento e della pipeline di pre-elaborazione.
Quando lo stato cambia, un messaggio viene inviato alla coda delle attività, per essere elaborato dal broker dei messaggi, che attiva il passaggio successivo della pipeline.

[3] http://www.celeryproject.org/

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR
SARDEGNA 2014-2020

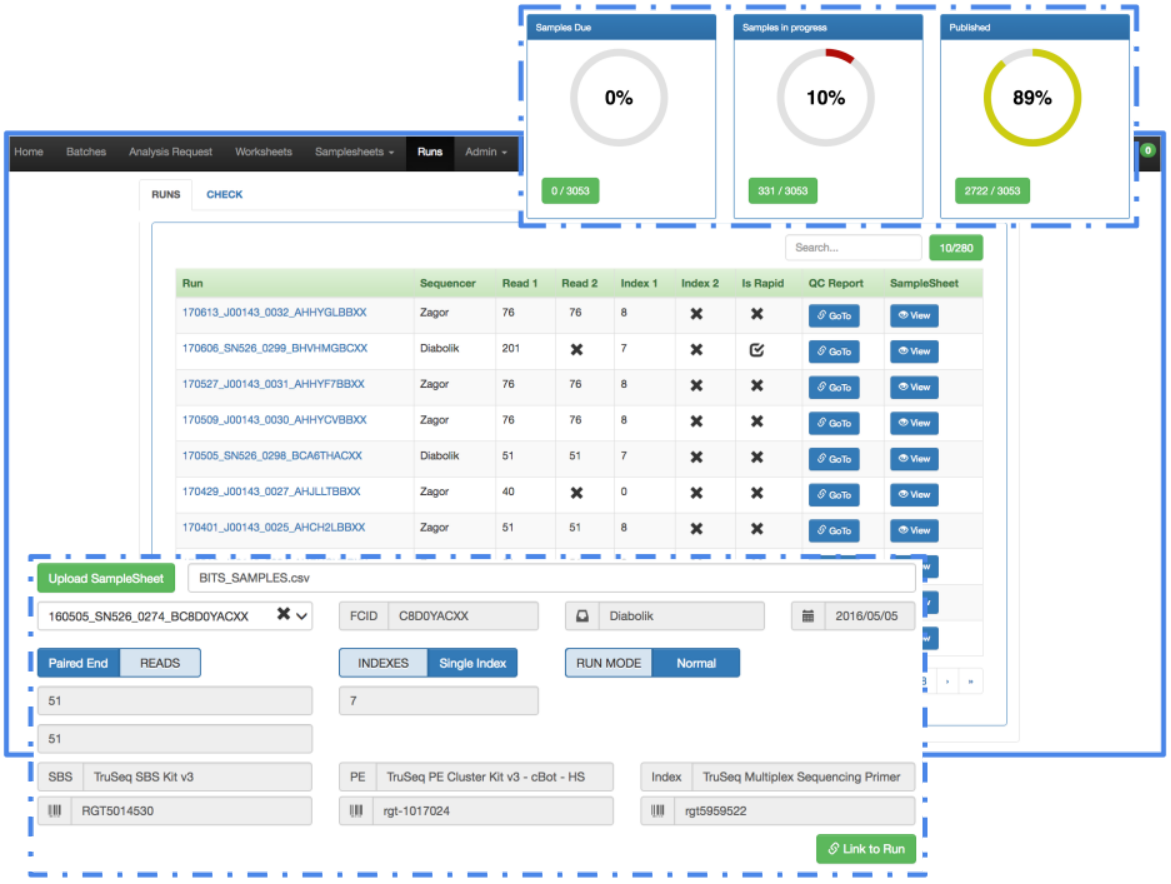

Fig. 3b - Screenshot di PENELOPE

Gestendo attività e messaggi, PRESTA è in grado di orchestrare automaticamente la conversione del formato, il demultiplexing, i controlli di qualità, la preparazione dei dati per l'archiviazione o la consegna.

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR
SARDEGNA 2014-2020

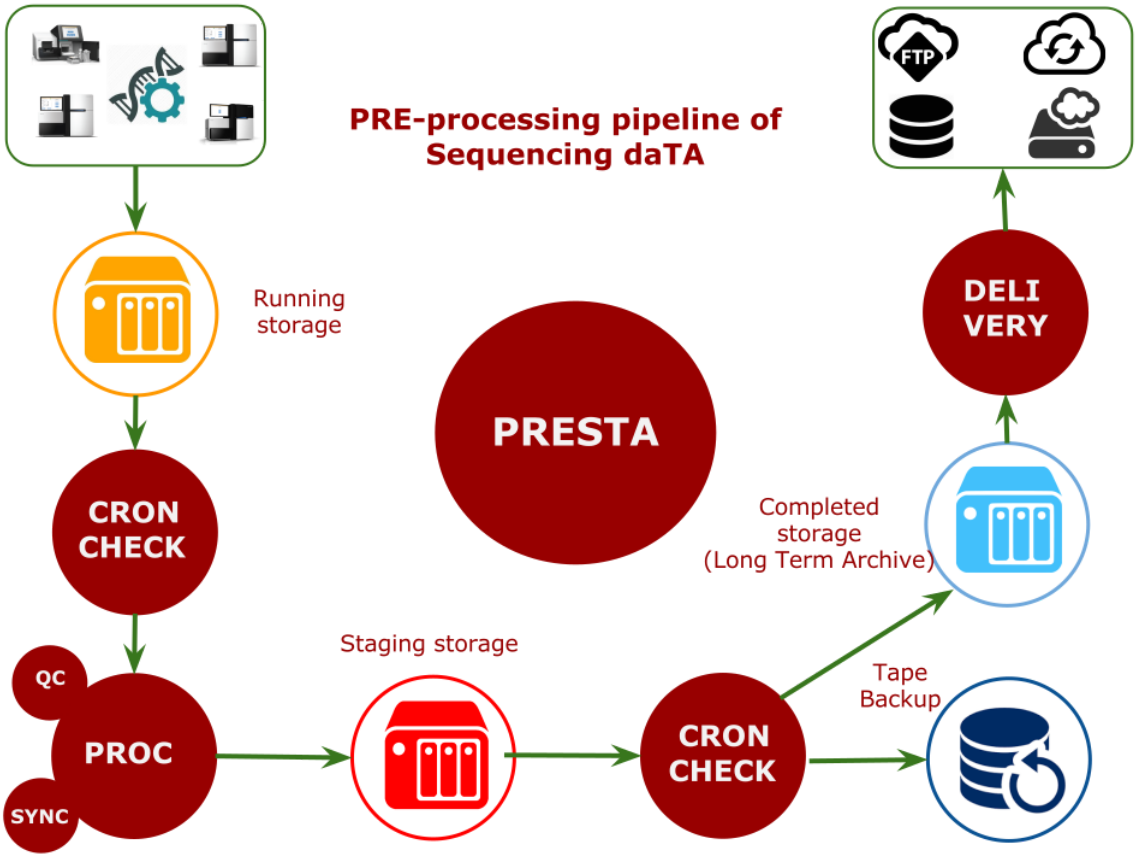

Fig. 4 - Schema della macchina a stati PRESTA

Gestione dei workflow di analisi computazionale

Per garantire la riproducibilità delle analisi bioinformatiche che verranno utilizzate nel progetto NIASMIC e la tracciabilità dei parametri utilizzati, le pipeline di analisi saranno implementate all'interno di un sistema di gestione di workflow.

A tale scopo, abbiamo sviluppato SOLIDA[4], un'applicazione Python da riga di comando che può facilmente organizzare l'implementazione, la gestione dei dati e l'esecuzione di un workflow basato su Snakemake[5], uno dei più potenti e diffusi sistemi di gestione workflow.

[4] https://github.com/gmauro/solida
[5] https://snakemake.readthedocs.io/en/stable/

Snakemake fornisce tutte le funzionalità necessarie per creare analisi di dati riproducibili e scalabili con una separazione logica tra la descrizione del flusso di lavoro e i dettagli di esecuzione.
SOLIDA può facilmente eseguire il bootstrap di qualsiasi workflow Snakemake organizzando la propria attività in diversi progetti correttamente configurati che possono essere diversi per il codice della pipeline, i dati di input, la configurazione del flusso di lavoro, l'ambiente virtuale e/o la cartella di lavoro.

Gestione dei dati fenotipici

Le informazioni associate ad ogni campione, dai dati fenotipici alla descrizione delle procedure computazionali utilizzate per l'analisi (*provenance*) saranno integrate in un database a grafo (Fig. 6). Questa tipologia di database, a differenza del modello relazionale classico basato su tabelle, grazie alla sua flessibilità è adatto alla gestione di dati eterogenei con caratteristiche che possono evolvere nel tempo.

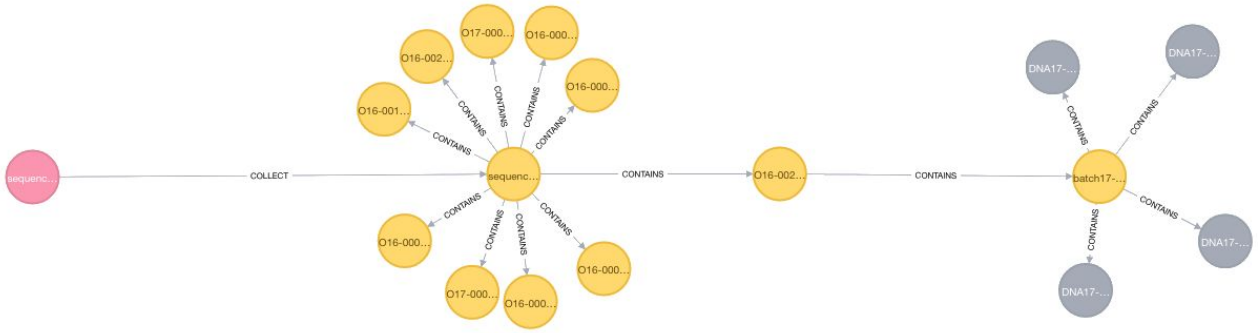

Fig. 6 - Schema di database a grafo, in cui le *connessioni* tra i *nodi* (ad es. campioni biologici, dati fenotipici, risultati di analisi) tracciano le informazioni sulla *provenance*.

Gestione dati di sequenziamento

Il sequenziamento NGS produce grandi moli di dati che devono essere organizzati in maniera efficiente per garantire l'accesso, anche su sistemi di storage fisicamente separati. A tale scopo utilizzeremo una tecnologia di *data virtualization* quale iRODS[6], che utilizza nomi logici univoci separati dai nomi con cui sono archiviati fisicamente, fornendo uno "spazio dei nomi logico" globale per tutti i set di dati.

[6] https://irods.org/

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR
SARDEGNA 2014-2020

iRODS può descrivere gli oggetti dati e le raccolte di dati tramite metadati ricchi e definiti dall'utente oltre ai metadati di sistema tradizionali. In questo modo, è disponibile un meccanismo di Data Discovery estremamente utile per individuare i dati rilevanti all'interno di set di dati di grandi dimensioni.

Pipeline di analisi

L'obiettivo dell'analisi della biopsia liquida è l'identificazione di mutazioni presenti nel DNA circolante che possano essere utilizzate per identificare i frammenti di DNA di origine tumorale e caratterizzare il profilo mutazionale del tumore. Nel progetto NIASMIC i dati ottenuti dal sequenziamento genomico del tumore primario, DNA circolante e buffy coat verranno analizzati con la suite di programmi GATK[7]. In particolare per le mutazioni somatiche verrà utilizzato il programma Mutect2[8], all'interno del framework GATK. Il software farà parte di una pipeline basata su Snakemake che includerà dei moduli per l'allineamento, il controllo di qualità pre- e post- allineamento, e per la ricerca di possibili contaminazioni nei campioni sequenziati.

[7] McKenna et. al 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome *Research 20:1297-303*
[8] Cibulskis et al. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 31, 213–219 (2013) doi:10.1038/nbt.2514

UNIONE EUROPEA
Fondo europeo di sviluppo regionale
REPUBBLICA ITALIANA
REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA
POR FESR SARDEGNA 2014-2020